Supplementary figures and tables

May 31, 2018

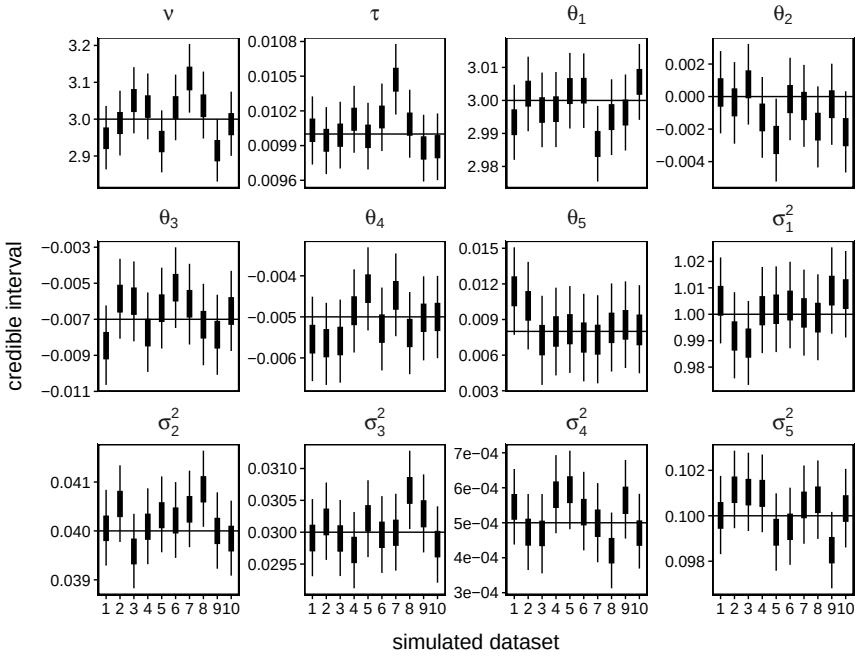

Figure S1: For Simulation Study 1 in Section 4.2, equal-tailed credible intervals for the hyperparameters, calculated from quantiles of MCMC samples. 50% credible intervals are shown as thick vertical lines, and 95% credible intervals are overlaid as narrow vertical lines.

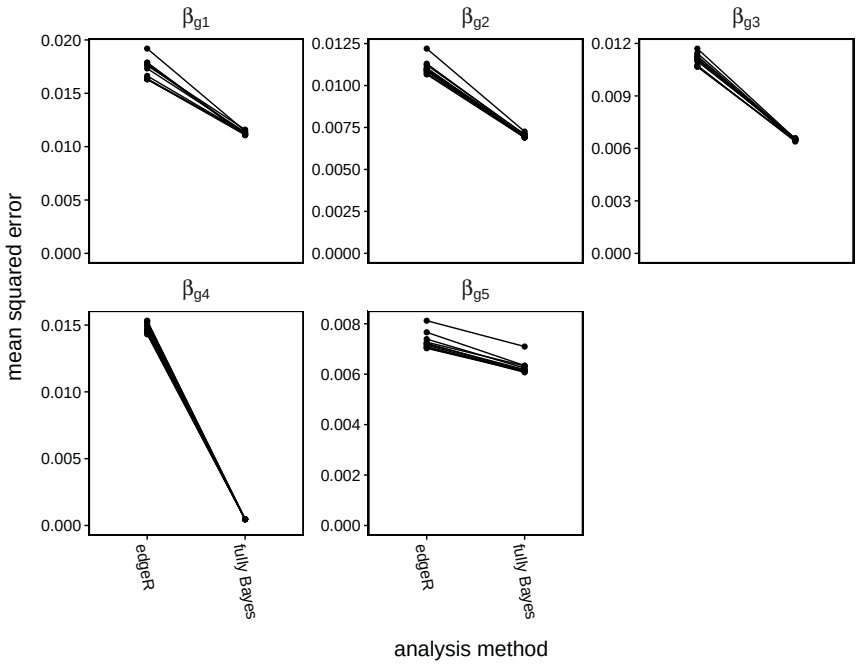

Figure S2: For Simulation Study 1 in Section 4.2, mean squared errors of the estimated model coefficients, where each mean is taken over all the genes. Each line corresponds to a simulated dataset.

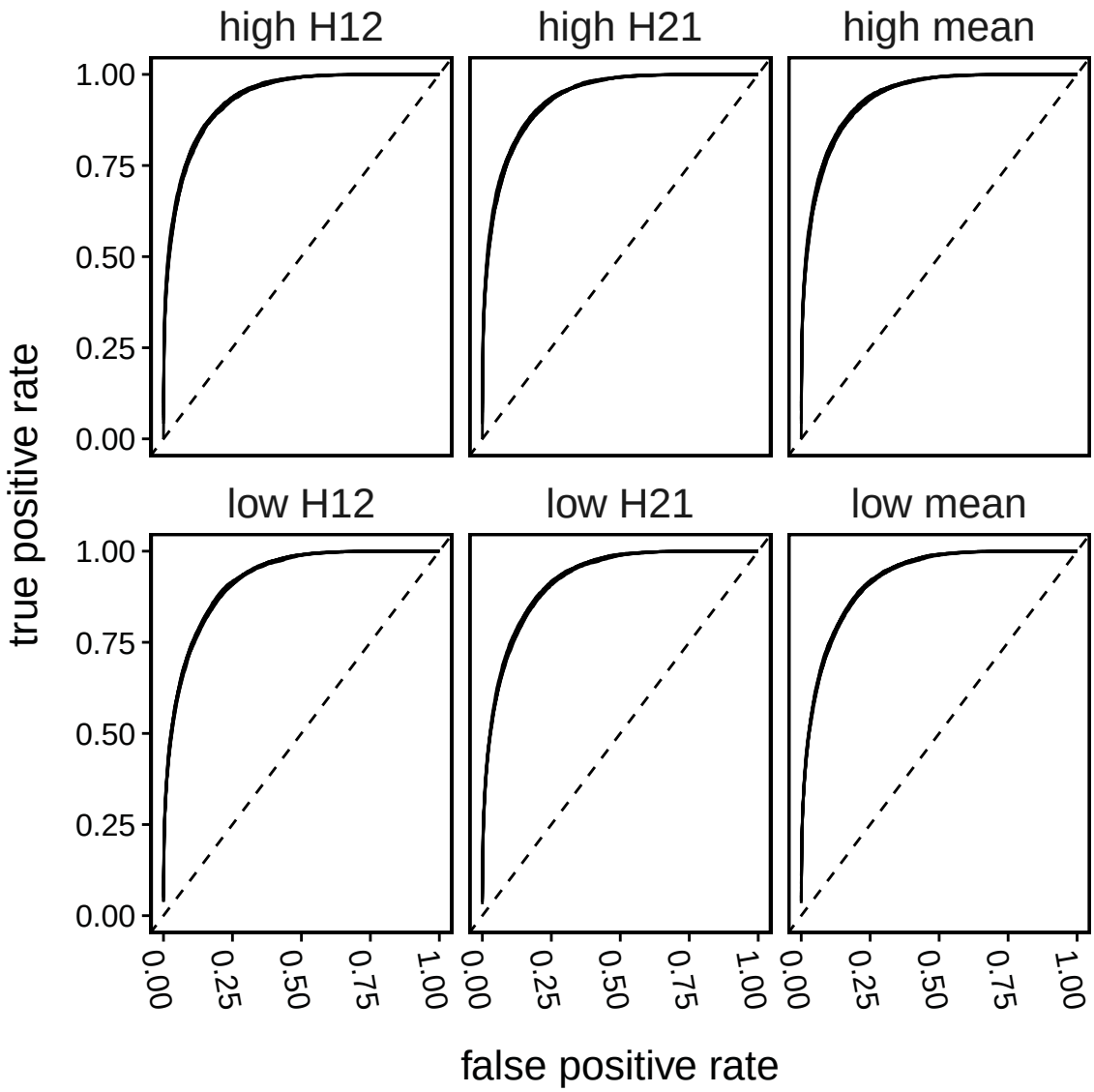

Figure S3: Receiver operating characteristic (ROC) curves for Simulation Study 1 in Section 4.2. There is one curve for each dataset and each kind of heterosis, and the dashed line is the identity line, the expected results of an ordering of genes completely at random. Areas under the curves range from 0.916 to 0.922 for low-parent heterosis and from 0.930 to 0.936 for high-parent heterosis.

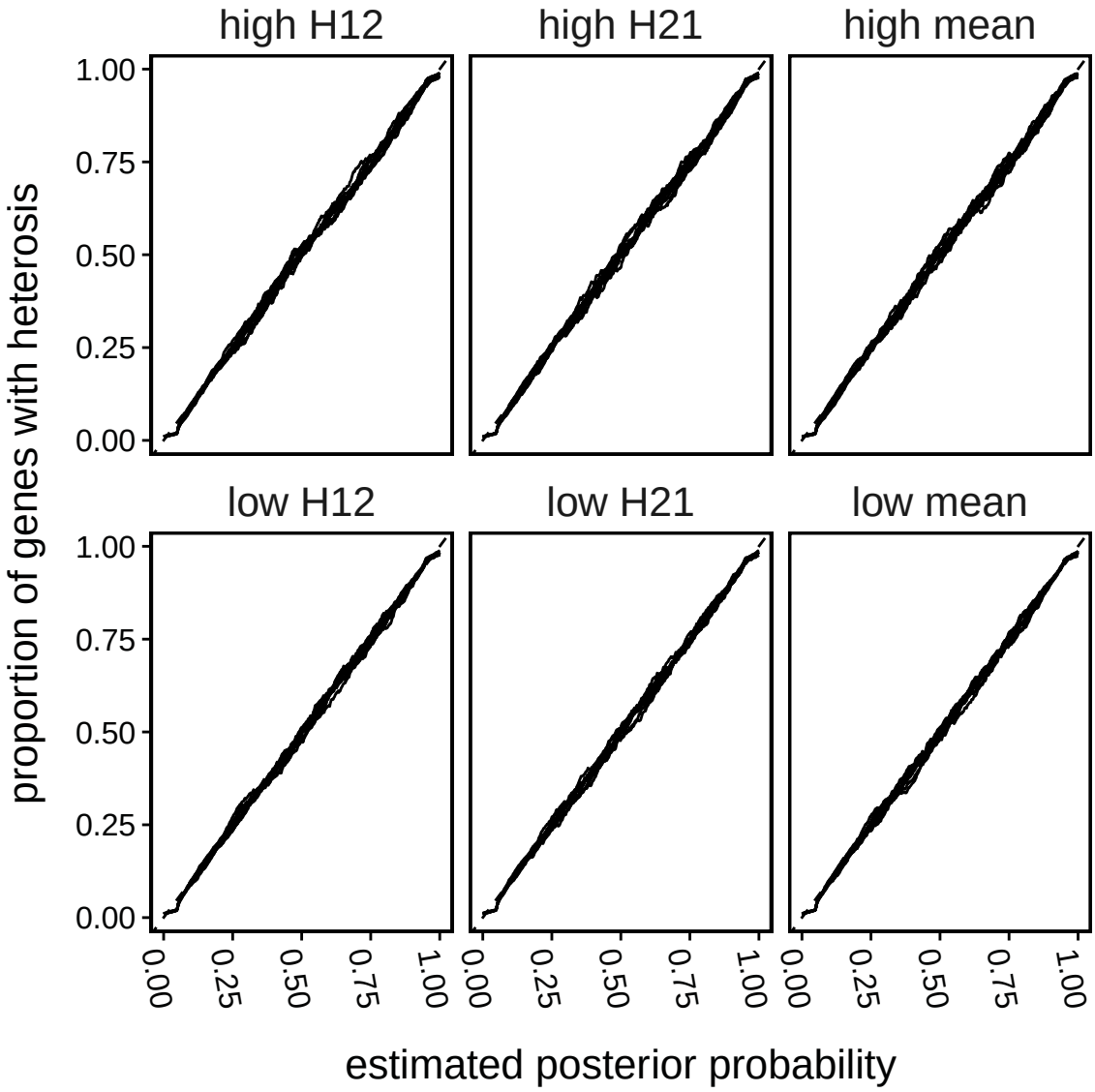

Figure S4: Calibration curves for Simulation Study 1 in Section 4.2. There is one curve for each dataset and each kind of heterosis. Each curve is the kernel-smoothed local proportion of true heterosis genes plotted against estimated probability from our fully Bayesian approach. The dashed line, hidden by the calibration curves, is identity line (the ideal calibration curve).

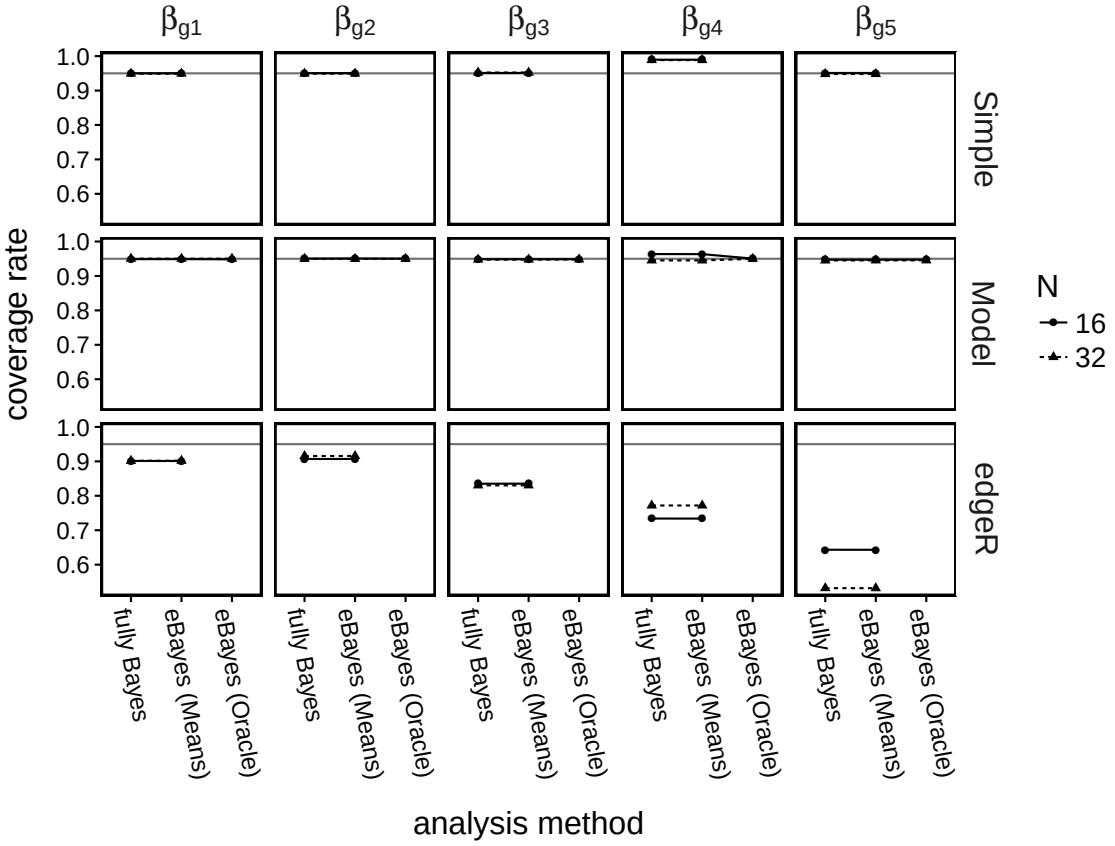

Figure S5: For Simulation Study 2 in Section 4.3, observed rates at which estimated 95% credible intervals cover parameters $\beta_{g\ell}$. The column labels indicate the $\beta_{g\ell}$ parameters, and the row labels indicate simulation scenarios. The gray horizontal lines indicate 0.95, the nominal coverage rate.

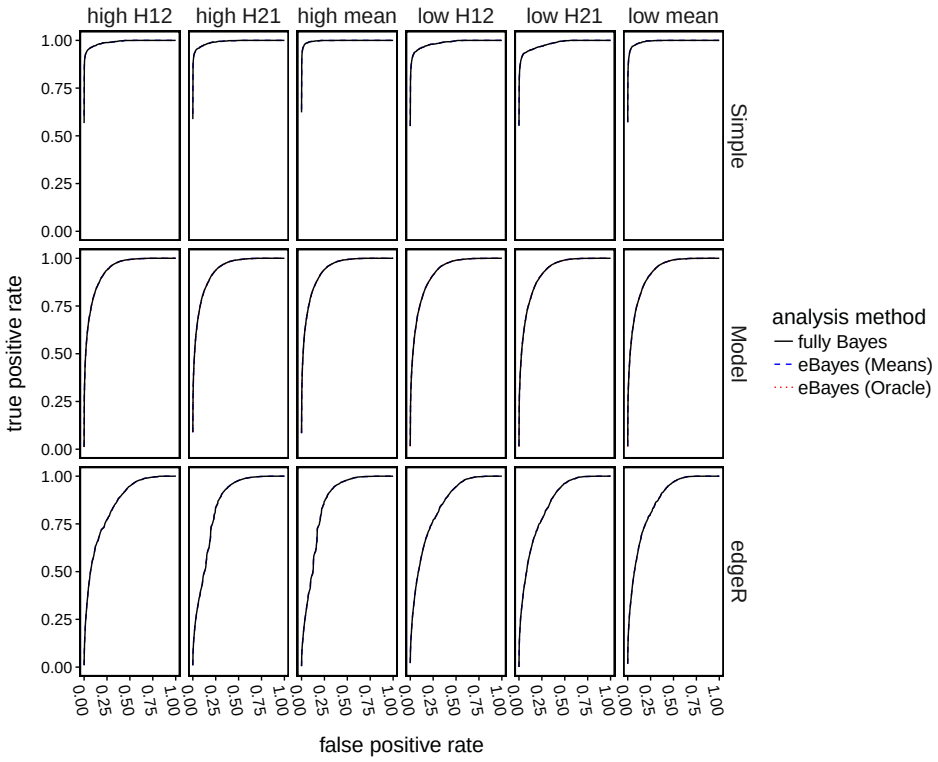

Figure S6: Receiver operating characteristic (ROC) curves for datasets with $N = 16$ in Simulation Study 2 in Section 4.3. The row labels indicate the method of simulating the data, and the column labels indicate the kind of heterosis detected. For heterosis, we use the notation from the general plant heterosis scenario from Section 2 and Table 1.

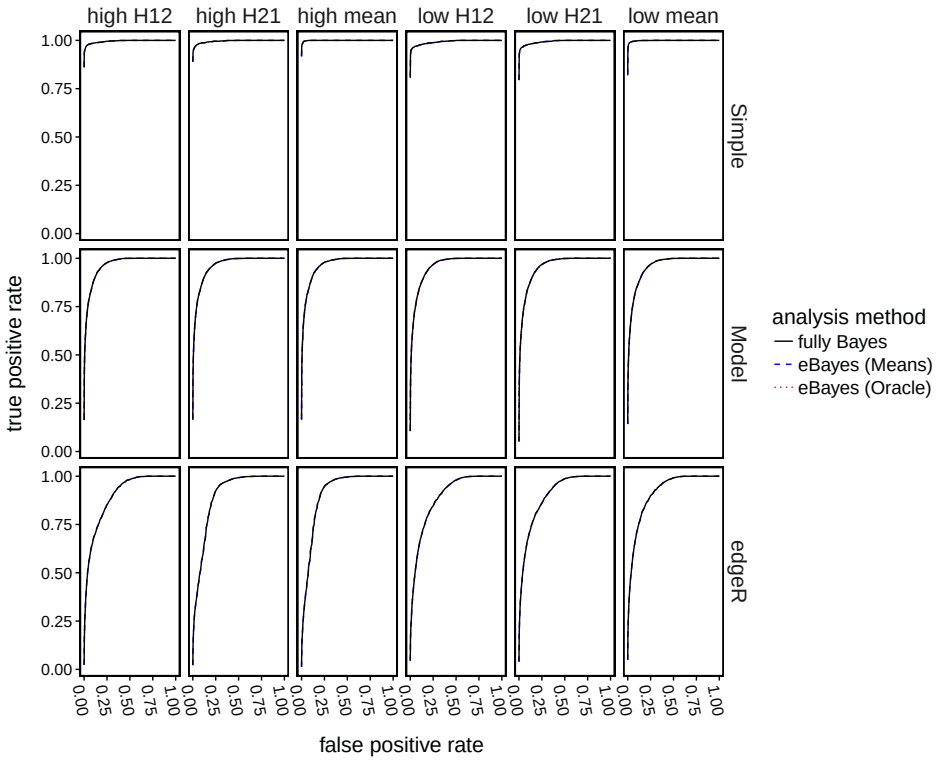

Figure S7: Receiver operating characteristic (ROC) curves for datasets with $N = 32$ in Simulation Study 2 in Section 4.3. The row labels indicate the method of simulating the data, and the column labels indicate the kind of heterosis detected. For heterosis, we use the notation from the general plant heterosis scenario from Section 2 and Table 1.

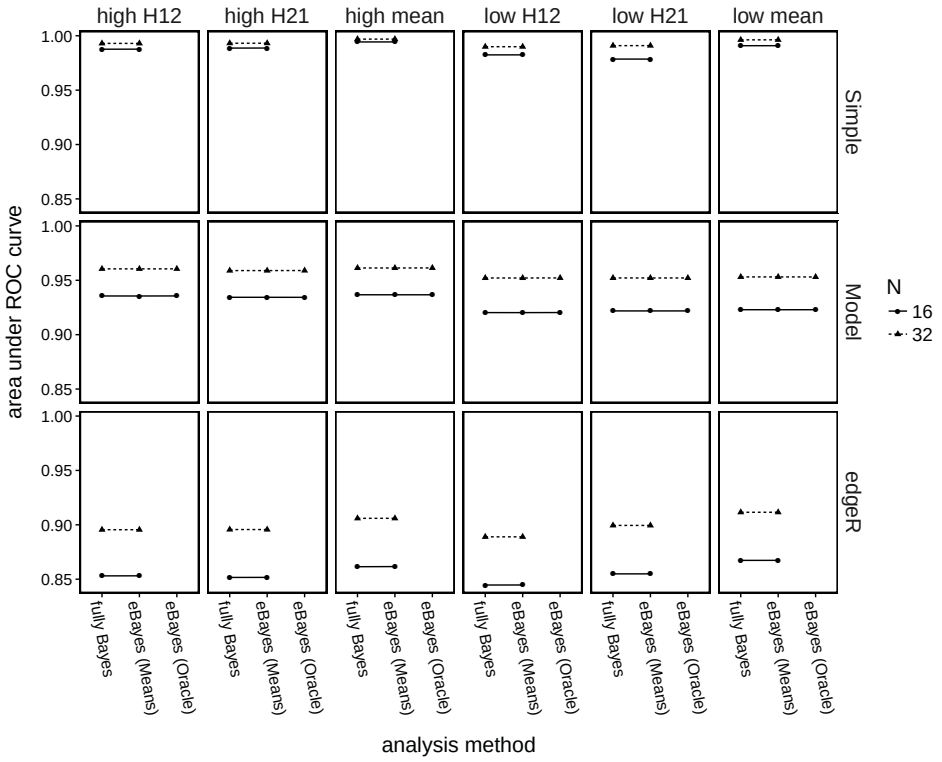

Figure S8: Areas under the receiver operating characteristic (ROC) curves for Simulation Study 2 in Section 4.3. The plotting shape denotes sample size ($N = 16$ or $N = 32$), the row labels indicate the method of simulating the data, the column labels indicate the kind of heterosis detected. For heterosis designations, we use the notation from the general plant breeding scenario from Section 2 and Table 1.

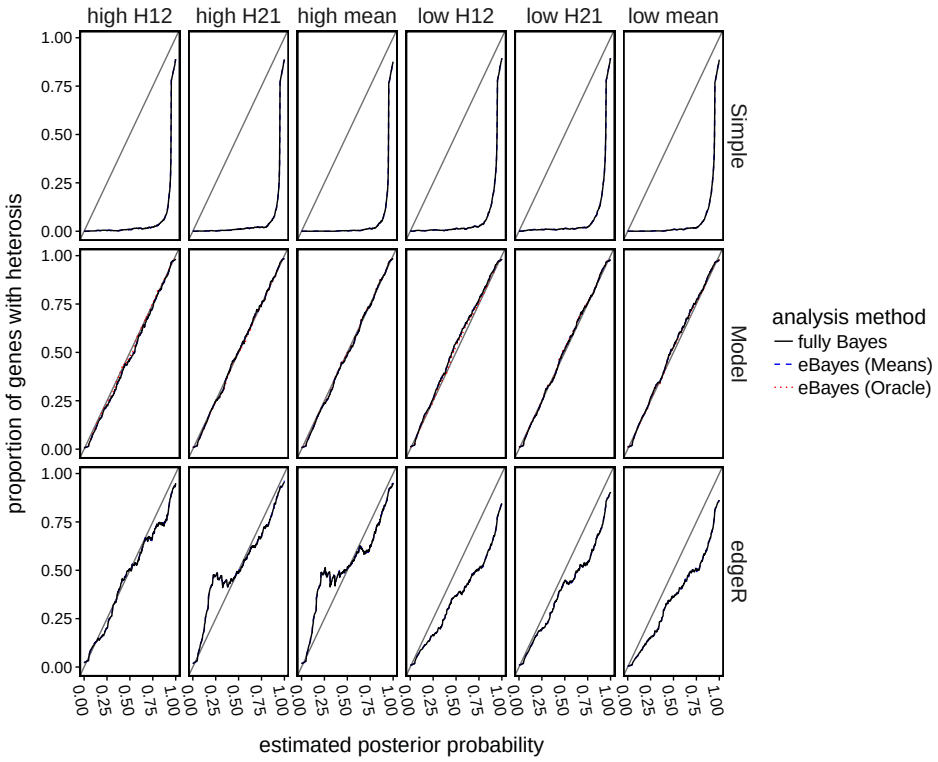

Figure S9: For datasets with $N = 16$ in Simulation Study 2 in Section 4.3, calibration curves for heterosis gene detection. The type of heterosis detected is indicated above each column, where we use the notation from the general plant breeding scenario from Section 2 and Table 1. The label to the right of each row indicates the method of simulating the data.

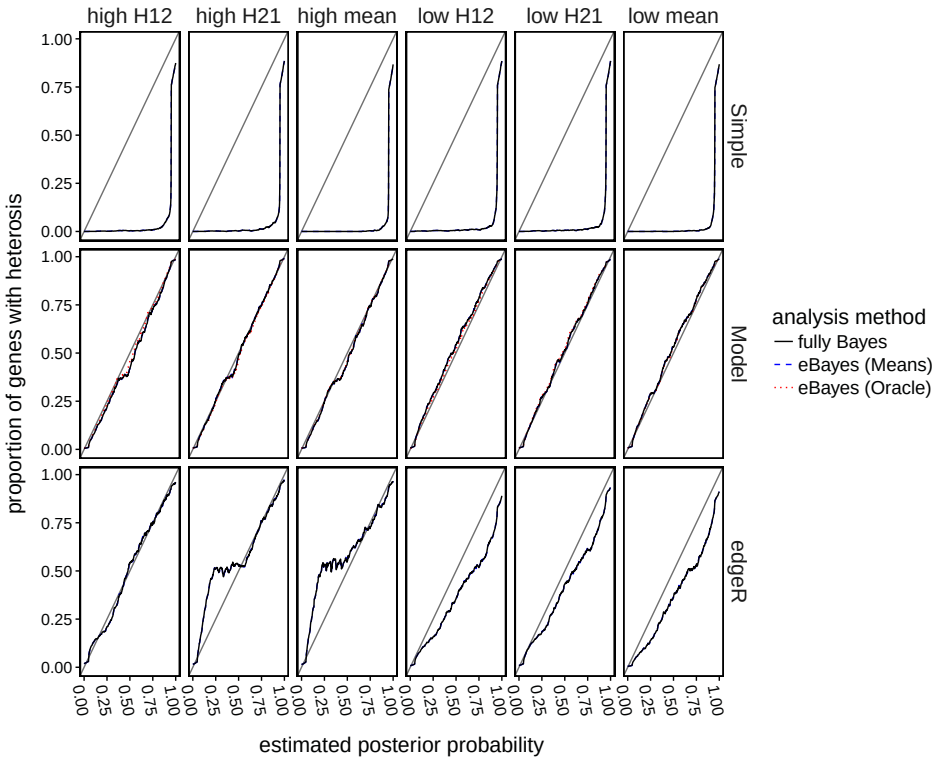

Figure S10: For datasets with $N = 32$ in Simulation Study 2 in Section 4.3, calibration curves for heterosis gene detection. A calibration curve, as explained in Section 4, is the smoothed local true proportion of heterosis genes plotted against posterior heterosis probability estimates from a statistical analysis. The identity line, plotted in solid gray in each panel, is the ideal calibration curve, which would result from perfectly accurate posterior probabilities. The type of heterosis detected is indicated above each column, where we use the notation from the general plant heterosis scenario from Section 2 and Table 1.The row labels indicate the method of simulating the data.

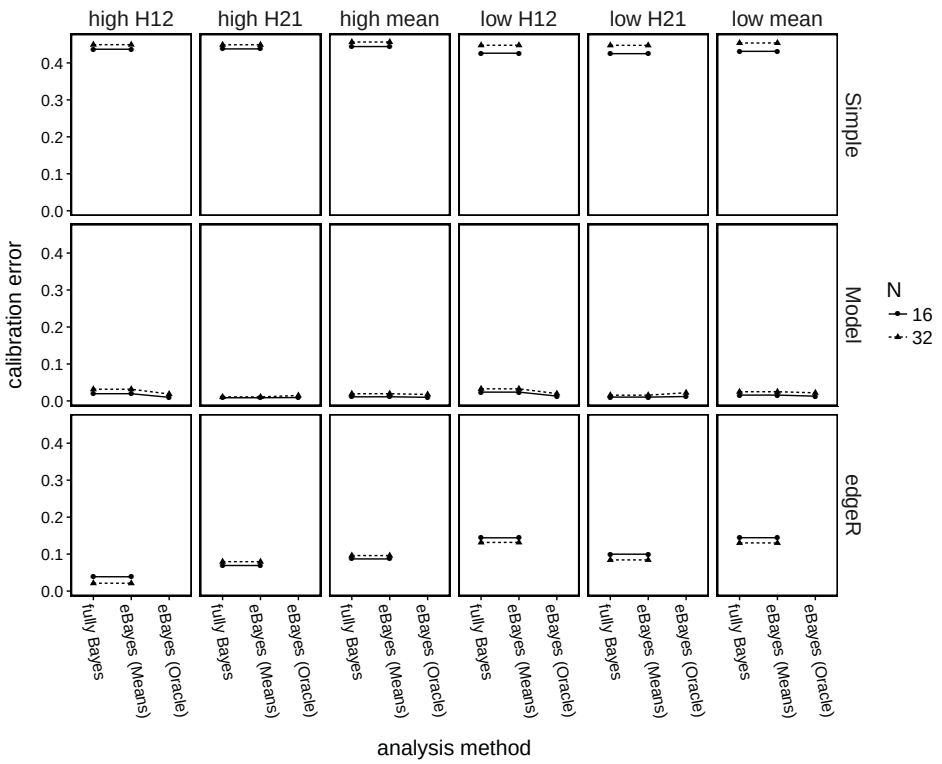

Figure S11: Mean absolute difference of each calibration curve in Figures S9 and S10 from the identity line. Triangle plotting symbols indicate simulated datasets with $N = 32$, and circles indicate datasets with $N = 16$. The type of heterosis is indicated above each column. For heterosis designations, we use the notation from the general plant breeding scenario from Section 2 and Table 1.

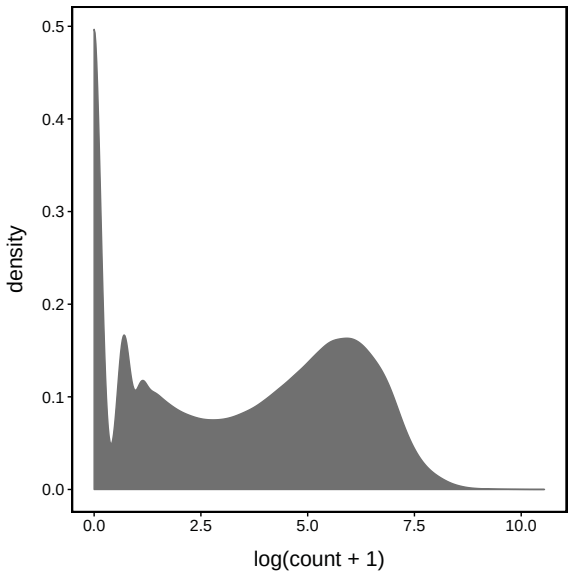

Figure S12: For the Paschold et al. dataset from Section 2, a kernel density estimate of the log of the counts after incrementing by 1.

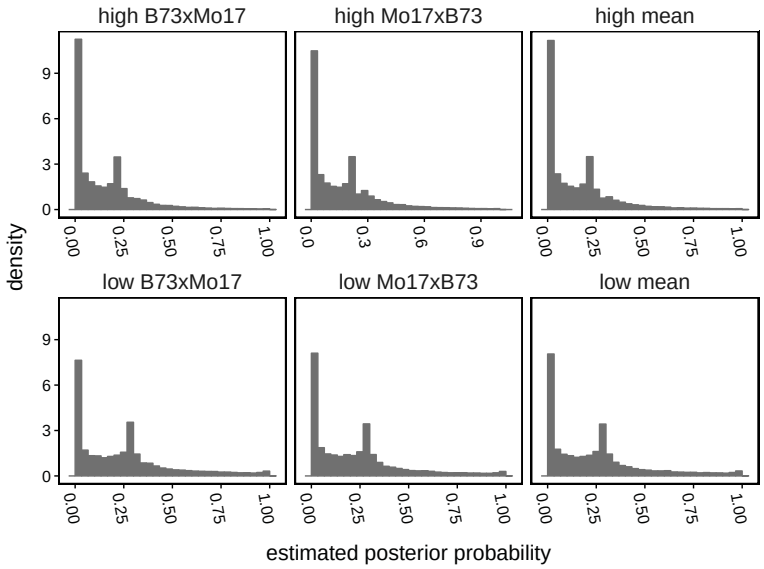

Figure S13: Histograms of estimated posterior probabilities of high (top row) and low (bottom row) heterosis for the B73×Mo17 hybrid (left column) and Mo17×B73 hybrid (middle column), and their mean (right column).

[Available separately as TableS1.csv]

Table S1: TableS1.csv is several megabytes in size and cannot be shown inline. This table, available for download, is a comma-separated values spreadsheet containing the total per-replicate gene expression counts of the Paschold et al. (2012) data, as well as posterior estimates of the gene-specific heterosis probabilities, the effect sizes, the means of the model coefficient parameters $\beta_{g\ell}$, hierarchical means γ_g, and the standard deviations of the $\beta_{g\ell}$'s and γ_g's from the fully Bayesian approach. The file also includes gene-specific parameter estimates from the edgeR method by McCarthy et al. (2012) from Section 4.1.

References

McCarthy, D., Chen, Y. and Smyth, G. (2012), 'Differential expression analysis of multi-factor RNA-seq experiments with respect to biological variation', *Nucleic Acids Research* **40**(10), 4288–4297.

Paschold, A., Jia, Y., Marcon, C., Lund, S., Larson, N. B., Yeh, C.-T., Ossowski, S., Lanz, C., Nettleton, D., Schnable, P. S. et al. (2012), 'Complementation contributes to transcriptome complexity in maize (*Zea mays L.*) hybrids relative to their inbred parents', *Genome research* **22**(12), 2445–2454.